رگرسیون درMinitab

امیرحسین محتسبی
amir@wizact.com

در نرم افزار مینی تب (ویرایش ۱۴) این امکان وجود دارد که براساس روش کوچکترین مربعات به محاسبه رگرسیون ساده (تک متغیره) و یا رگرسیون چند متغیره پرداخت. رگرسیون نوع ساده فقط براساس برآورد کوچکترین مربعات یک سری از داده ها تخمین مورد نظر را محاسبه می کند اما در رگرسیون چند متغیره این تخمین بین یک عامل مورد نظر برای پیش بینی , مانند وزن , و دو یا چند پیش بینی کننده , مانند قد و سن , مطرح است. همانطور که گفته شد در رگرسیون اغلب از روش کمترین مربعات استفاده می شود و معادله خطی را تعیین می کند که دارای کمینه مجموع مربعات فواصل افقی بین نقاط داده و خط مورد نظر باشد.

در بحث رگرسیون به دو مفهوم زیاد برخورد می کنیم :

- Response , همان ستون محتوی اقلام **Y** و یا متغیر جواب.
- Predictores , ستون یا ستونهای **X** و یا تخمین زننده ها.

در هنگام محاسبه رگرسیون باید یک ستون برای متغیر جواب و حداقل یک ستون برای تخمین زننده ها داشته باشیم. این ستون ها باید دارای طول یکسان باشند. همچنین ردیف هایی که دارای مقدار تهی باشند از محاسبه حذف خواهند شد. نکته دیگر این است که اگر مقدارهایی بسیار به هم نزدیک باشند و یا حالت تقریبا ثابت داشته باشند همگی و یا تعدادی از آنها از محاسبه حذف خواهند شد[۱].

برای انجام یک رگرسیون چند متغیره مثال زیر را دنبال می کنیم :

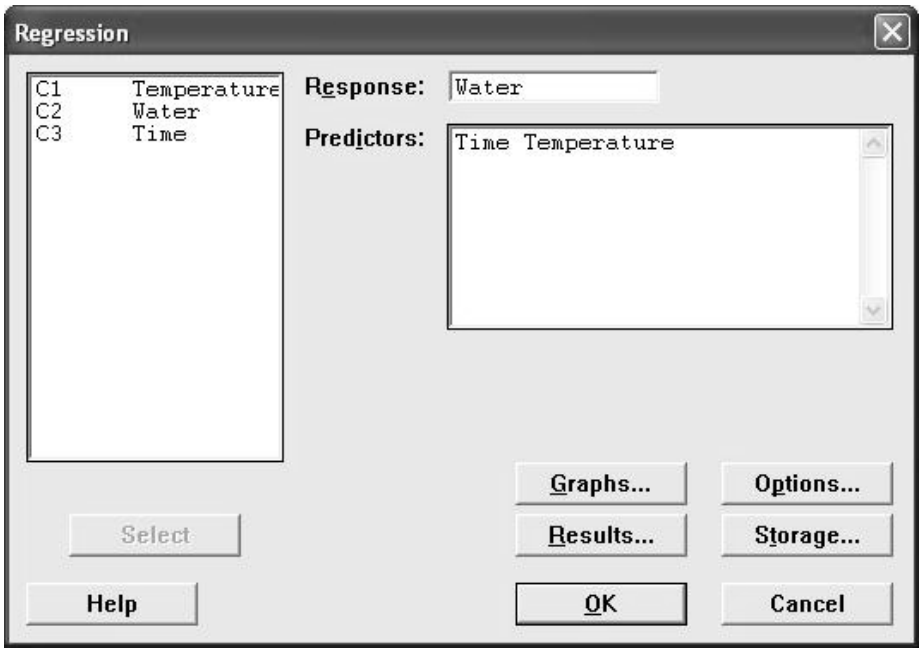
[۱] به این حالت **ill-conditioned** می گویند و می توانید در نرم افزار شرایطی که منجر به این تشخیص می شود را تنظیم کنید.

مثال : فرض کنید که در یک بازه زمانی سه ساعته سه متغیر دما (Temprature) زمان مصرف شده برای زدن چمن ها (Time mowing the grass) و مصرف آب (Water Consumption) را اندازه گیری کرده ایم و جدول زیر را بدست آورده ایم.

Temperature (F)	Water Consumption (ounces)	Time mowing the grass (hours)
۷۵	۱۶	۱,۸۵
۸۳	۲۰	۱,۲۵
۸۵	۲۵	۱,۵
۸۵	۲۷	۱,۷۵
۹۲	۳۲	۱,۱۵
۹۷	٤۸	۱,۷۵
۹۹	٤۸	۱,٦

ابتدا یک صفحه کاری جدید ایجاد کرده و اطلاعات را در آن وارد می کنیم...

از منوی Stats>Regression>Regression وارد قسمت رگرسیون می شویم و به عنوان Response ستون مصرف آب و به عنوان Predictores ستون زمان زدن چمن و دما را انتخاب می کنیم. و کلید Ok را بزنید.



اگر همه چیز درست پیش رفته باشد نتیجه ای مشابه زیر در پنجره خط فرمان خواهید دید:

```
The regression equation is
Water = - ۱۲۲ + ۱۲,۵ Time + ۱,۵۱ Temperature


Predictor        Coef  SE Coef       T      P
Constant     -۱۲۱,٦٥٥    ٦,٥٤٠  -۱۸,٦۰  ۰,۰۰۰
Time          ۱۲,٥۳۲     ۱,۹۳۳    ٦,٤۸  ۰,۰۰۳
Temperature   ۱,٥۱۲۳٦  ۰,۰٦۰۷۷   ۲٤,۸۹  ۰,۰۰۰


S = ۱,۲٤٤۸۸   R-Sq = ۹۹,٤٪  R -Sq(adj) = ۹۹,۰٪


Analysis of Variance

Source          DF      SS      MS       F      P
Regression       ۲  ۹۷۰,٦٦  ٤۸٥,۳۳  ۳۱۳,۱۷  ۰,۰۰۰
Residual Error   ٤    ٦,۲۰    ۱,٥٥
Total            ٦  ۹۷٦,۸٦


Source       DF  Seq SS
Time          ۱   ۱۰,۸۸
Temperature   ۱  ۹٥۹,۷۸
```

تفسیر نتیجه:

عدد ثابت **ما برابر** ۱۲۱,٦٥٥- **است.این به این معنا است که انتظار می رود کارگر حدود ۱۲۱ اونس آب مصرف می کند در حالتی که دما برابر صفر باشد و هیچ زمانی هم برای چمن زنی مصرف نکند. بنابر این تابع ما برای** x=۰ **درست عمل نمی کند و فرض ما یک روز تابستانی با دمای حدود ۷۵ تا ۹۹ درجه فارنهایت می باشد.**

دما **: به ضریب زاویه خط مربوط می شود و حدوداً برابر ۱,۵ است. این ضریب زاویه برابر تقسیم میزان اونس آب مصرفی به دما بر حسب فارنهایت می باشد. یعنی به ازای هر درجه افزایش در دما انتظار می رود مصرف آب حدود ۱,۵ اونس افزایش یابد اگر متغیر زمان ثابت فرض شود.**

زمان زدن چمن : به همین ترتیب باز هم ضریب زاویه خط است و برابر تقسیم میزان اونس آب مصرفی به زمان زدن چمن بر حسب زمان است و حدوداً برابر ۱۲,۵ است. این عدد مشخص می کند که به ازای افزایش هر ساعت زدن چمن ها میزان اونس آب مصرفی حدود ۱۲,۵ واحد افزایش می یابد.

در مثال بالا ابتدا معادله رگرسیون را می بینید :

```
The regression equation is
Water = - ۱۲۲ + ۱۲,۵ Time + ۱,۵۱ Temperature
```

سپس در جدول هر یک از تخمین زننده ها آمده اند :

Predictor	Coef	SE Coef	T	P
Constant	-۱۲۱,٦٥٥	٦,٥٤٠	-۱۸,٦۰	۰,۰۰۰
Time	۱۲,٥۳۲	۱,۹۳۳	٦,٤۸	۰,۰۰۳
Temperature	۱,٥۱۲۳٦	۰,۰٦۰۷۷	۲٤,۸۹	۰,۰۰۰

ضریب هر یک از متغیر ها در ستون Coef[۲] مشخص شده اند. خطای استاندارد هر یک از تخمین ها را نیز در ستون SE Coef[۳] می بینیم. برای بدست آوردن فاصله اطمینان ٪۹۵ این عدد را در ۱,۹٦ ضرب کرده و آنرا از ضرایب اضافه و کم می کنیم. به عنوان مثال برای دما داریم :

```
۱,٥  + ۰,۰٦ * ۱,۹٦=۱,۳۸۲٤
۱,٥  -  ۰,۰٦ * ۱,۹٦=۱,٦۱۷٦
```

در این حالت مقدار صحیح این ضریب به احتمال بسیار زیاد در بین دو عدد ۱,۳۸۲٤ و ۱,٦۱۷٦ است و فقط ٪٥ احتمال دارد که این گفته مان اشتباه باشد.

ستون بعدی T است که برابر است با :

T=(Coef/SE)

T به تنهایی زیاد به کار نمی آید اما برای محاسبه P از آن استفاده می شود. P در آزمون فرض به ما کمک می کند که جواب را قبول و یا رد کنیم. در اصل P احتمال رخداد خطای نوع اول است. اگر مانند Temprature و Water این احتمال کمتر از ۰,۰۰۰٥ باشد (۰,۰۰۰) به این معنی است که رابطه[۴] بین این دو از لحاظ آماری برای α=۰,۰٥ بسیار قابل توجه است.

---

[۲] coefficients
[۳] standard error
[٤] Relation

همانطور که ملاحظه می شود مقدار P برای Temperature برابر صفر است اما این مقدار برای Time برابر ۰,۰۰۳ است و به این معنی است که Temperature به شدت با Water در ارتباط است . در نتیجه مدل ما با Temperature به تنهایی بیشتر برازنده است.

در حال حاضر ما یک تخمین از میزان مصرف آب و دما و زمان زدن چمن ها داریم. اما این تخمین چقدر مورد اطمینان است؟ میزان r-sq برابر ٪۹۹,٤ است. یعنی دما حدود ٪۹۹,٥ مواقع میزان آب مصرفی را توضیح می دهد که تخمین خیلی بالایی است. اگر ما این عدد را به صورت ۰,۹۹۵ در نظر بگیریم و ریشه مربع آنرا بیابیم عدد ۰,۹۹۷ را خواهیم داشت که برابر همبستگی بین میزان پیش بینی شده و میزان واقعی مصرف است.

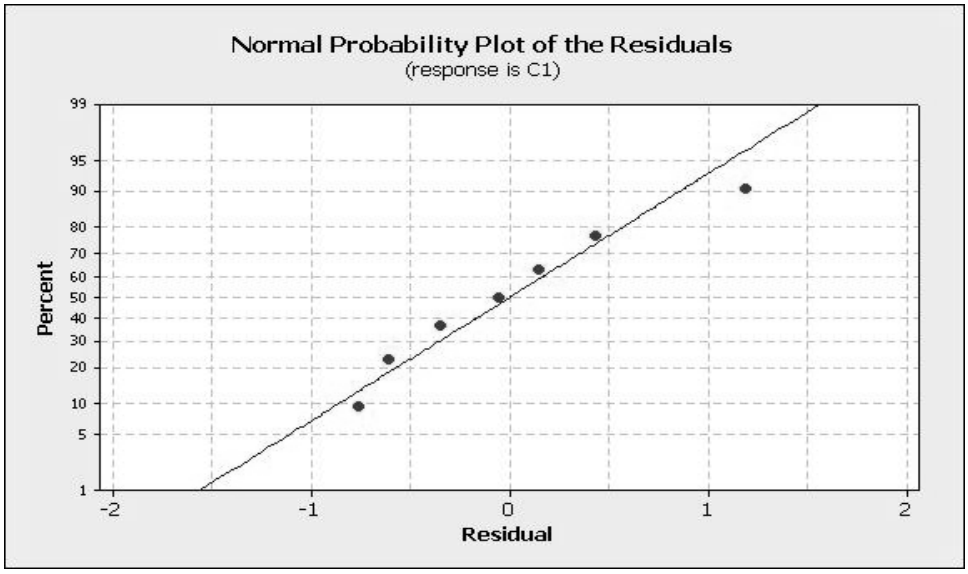
اگر در صفحه تنظیم رگرسیون بر روی دکمه گراف کلیک کرده باشید این امکان را دارید که از نمودارهایی که برای شما رسم می شود جهت تحلیل بهتر داده ها استفاده کنید. انواع این نمودار ها شامل :

- **Histogram of Residuals**

هیستوگرام, الگویی را بر اساس توزیع نرمال نمایش می دهد و بیشتر برای نمونه داده هایی با جمعیت بیشتر از ۵۰ عدد کارایی دارد.

- **Normal Probability Plot**

نمودار احتمال نرمال یک الگوی خطی مبتنی بر توزیع نرمال است .



مراجع:

- **Using Minitab, http://mtsu۳۲.mtsu.edu:۱۱۳۰۸/index.html**
- **Minitab Software Manual**
- **Google Answers http://answers.google.com**

این مقاله را در اینجا بیابید :

http://www.wizact.com/articles/articles.aspx?id=۲۷